COMPSCI 752 Web Data Management and Distribution

Course Outline

THIS COURSE IS MANAGED WITH CECIL. PLEASE REFER TO CECIL FOR CURRENT INFORMATION AND COURSE CONTENTS.

Schedule

First Semester 2015, for current timetable and rooms please refer to university timetabling system; please watch out for room changes in the first week of semester.

Assessment

Implementation 10%, Presentation 10%, Assignments 10%, Report 25%, Demo 5%, Exam 40%

Description

The Internet and the Web have revolutionized our access to information. Today information on the Web is mostly HTML as well as pdf, doc, plain text and also images, music and videos. The public Web is composed of billions of pages on millions of servers. It is a fantastic means of sharing information. It is very simple to use for humans. On the negative side, it is very inappropriate for access by software applications. This motivated the introduction of new data models, namely XML and RDF that are well suited both for humans and machines. The course aims to describe the structure of information found on the Web, and to explain how this information can be efficiently represented, described and accessed. Primary topics of the course include Web data modelling and large-scale data management in distributed and heterogeneous environments.

Contents

Extensible Markup Language (XML), XPath, XQuery, XUpdate, XSLT, Tree automata, RDF, RDFS, Ontologies, OWL, SPARQL, Data Integration, Web Search, Data distribution, Distributed computing, Hadoop, MapReduce, Pig

Coursework

- The implementation (10%) is due 24 April 2015, 5pm
- Report (25%) is due 1st May 2015, 5pm
- Assignment 1 (5%), due Monday 30 March, 5pm
- Assignment 2 (5%), due 12 May, 5pm
- Presentation (10%) and Demo (5%), in the second half of the semester.

Recommended Textbook

- *Web Data Management*. Serge Abiteboul, Ioana Manolescu, Philippe Rigaux, Marie-Christine Rousset, Pierre Senellart; Cambridge University Press, 2011
  http://webdam.inria.fr/Jorge/

Related Reading

Library and web resources will be indicated, along with some handouts in lectures.

Lecturer

- Dr Gerald Weber – course director

Recommended Prerequisite

COMPSCI 351 or COMPSCI 751

Learning Outcomes

Students will be able to:

- Apply the state of the art in representation formalisms for Web data, including the eXtensible Markup Language (XML) and the Resource Description Framework (RDF)
- Model Web data with XML schema languages, including Document Type Definitions (DTDs), XML Schema and tree automata
- Query Web data with XML and RDF query languages, including XPath, XQuery, XSLT, and SPARQL
- Integrate Web data with ontologies, including RDF schema and the Web Ontology Language OWL
- Understand how to manage big data on the Web, including techniques for searching, indexing and processing such as PageRank, BigTable and MapReduce